Single-nucleotide polymorphisms in the public domain: how useful are they?

There is a concerted effort by a number of public and private groups to identify a large set of human single-nucleotide polymorphisms[1,2] (SNPs). As of March 2001, 2.84 million SNPs have been deposited in the public database, dbSNP, at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/SNP/). The 2.84 million SNPs can be grouped into 1.65 million non-redundant SNPs. As part of the International SNP Map Working Group, we recently published a high-density SNP map of the human genome consisting of 1.42 million SNPs (ref. 3). In addition, numerous SNPs are maintained in proprietary databases. Our survey of more than 1,200 SNPs indicates that more than 80% of TSC and Washington University candidate SNPs are polymorphic and that approximately 50% of the candidate SNPs from these two sources are common SNPs (with minor allele frequency of ≥20%) in any given population.

Most of the SNPs in the public domain came from three groups: the SNP Consortium[4] (TSC), the Sanger Centre in the United Kingdom, and Washington University[5]. The SNPs found in dbSNP are mostly 'candidate' SNPs found by computer data-mining procedures and have not been characterized. In other words, the SNPs in dbSNP are mostly variants found when DNA sequences from a handful of clones were compared by a computer algorithm[6]. They are basically annotations of the human genome sequence. By our estimate, less than 15% of the SNPs in the database have been proven to be polymorphic in any population. Even fewer have genotyping assays developed for them. We carried out two pilot studies to determine how well the candidate SNPs in dbSNP would fare if they were to be developed into genetic markers.

In the first study, 528 radiation hybrid mapped sequence-tagged sites (STSs) containing candidate SNPs from the TSC set were tested by a pooled DNA sequencing approach to determine the allele frequencies of the SNPs in 3 ethnic groups[7] (Caucasians, Chinese and Africans). Each DNA pool contains equal amounts of DNA from 30 individuals. Preparative PCR (30 μl reactions) were carried out in 96-well microtiter plates and the excess PCR primers and deoxynucleotides were removed by passing the crude PCR products through a size-exclusion resin in 96-well format (Edge Biosystems). An aliquot of the PCR product was used in the sequencing reaction (also done in 96-well format) and the sequencing reaction products were purified by a size-exclusion resin (Princeton Separations). We found that 28 STSs failed PCR and sequencing (5.3%; Table 1). In the 500 STSs that amplified well, we found 539 candidate SNPs. Complete sequencing data were obtained for 502 candidate SNPs (93%). The remaining 7% of candidate SNPs only had partial data (that is, one or more of the pool sequences were missing). Of the characterized SNPs, 87 (17%) were monomorphic (that is, only 1 of the 2 predicted alleles was found in all 3 population samples), and 30 SNPs (6%) had minor allele frequencies below 20% in all 3 populations. In contrast, 135 SNPs (27%) were common SNPs, with minor allele frequencies greater than or equal to 20% in all 3 population samples; 263 SNPs (52%) were common in 2 or more populations; and 385 (77%) were common SNPs in at least 1 population.

In a second study, STSs were developed for 897 candidate SNPs generated by comparing the consensus genomic sequences of two overlapping BAC clones. The STSs were developed using the Primer3 program[8]. In all, 133 STSs failed PCR and sequencing (14.8%), leaving 774 candidate SNPs found in 764 STSs. Similar to the results obtained in the TSC pilot project, 130 candidate SNPs (16.8%) were monomorphic. We found 55 SNPs (7.1%) to have a minor allele frequency of less than 20% in all 3 population samples; 208 (26.9%) to be common SNPs in all 3 population samples; 420 (54.3%) to be common SNPs in 2 or more populations; and 589 (76.1%) to be common SNPs in at least 1 population.

In both studies, between 52% and 54% of the characterized SNPs turn out to be common SNPs for each population pool. In other words, about half of the candidates are common SNPs in the Caucasians, and so forth. Moreover, between 30% and 34% of the characterized SNPs are not detected in each population pool. Our results show that if a researcher uses the publicly available candidate SNPs for a study in a population, there is only a 66–70% chance that the SNPs have appreciable minor allele frequency and a 50-50 chance that the SNPs are common in that population.

Although pooled sequencing for allele frequency estimation is a validated method[9], and our recent study showed that the individual genotype data (over 300 individuals typed for each marker) corresponded well with the pooled sequencing data[10], this approach yields only a rough estimate of the allele frequencies. Because of uncertainties in the accuracies of DNA pooling and sequencing data quality issues, one cannot detect rare alleles (<5% in the pooled sample) and the allele frequency estimates can be off by about 5% (ref. 9). Here we are only trying to determine if a candidate SNP has appreciable minor allele frequency and if it is a common SNP in a population. Both questions can be answered with confidence in this rough estimation approach.

There is also a real concern that the candidate SNPs are not real polymorphisms but duplicated regions of the genome with near-identical sequences. This is a legitimate concern, except that the candidate SNPs from the TSC we used were uniquely mapped by radiation hybrid mapping and the overlap SNPs were from clone sequences with extensive alignment in the vicinity. With the increasing complete human genome sequence as reference, most of the false-positive SNPs due to paralogous sequences have already been screened out as the SNPs are mapped. Moreover, the

Table 1 • Allele frequencies of SNPs found in dbSNP

	TSC SNPs		Overlap SNPs	
Total characterized	502		774	
SNPs not detected[a]	87	(17.3%)	130	(16.8%)
Uncommon SNPs[b]	30	(6.0%)	55	(7.1%)
Common SNPs in ≥1 population[c]	385	(76.7%)	589	(76.1%)
Common SNPs in ≥2 populations[c]	263	(52.4%)	420	(54.3%)
Common SNPs in all 3 populations[c]	135	(27.0%)	208	(26.9%)

[a]Only one of the two predicted alleles found in all three populations. [b]Minor allele frequency appreciable but <20% in all 3 populations. [c]A SNP is considered 'common' when the minor allele frequency is ≥20%.

false-positive SNPs in duplicated regions show the tell-tale sign of having 50% allele frequencies for both alleles in all populations. The only way to test for false-positive SNPs due to duplications is to check for mendelian inheritance of the alleles or assay the candidate SNP against a duplicated haploid genome such as the complete hydatidiform mole[5]. Based on the general experience that only approximately 5% of candidate SNPs that passed the computer filters for repetitive elements are due to low-copy duplications, global testing of candidate SNPs for duplications is not warranted.

Because a significant fraction of the SNPs in the public domain are found in repetitive regions, there is no guarantee that all SNPs can be amplified uniquely from the genome. Despite these limitations, the publicly available candidate SNPs from TSC and Washington University are likely to be useful to any researcher looking for SNPs in the public domain if they are selected judiciously. To make the marker set even more useful to the genome research community, our group at Washington University and several other groups will characterize more than 100,000 candidate SNPs by the end of 2001. With PCR assays designed for the SNPs and the allele frequencies of these SNPs determined, the average researcher can use these SNPs with a high degree of confidence that they are useful in their own populations.

Acknowledgments

We thank E.P.H. Yap for the Asian samples; M. Boyce-Jacino for the Caucasian and African American samples; R. Sachidanandam and L. Stein for TSC sequence information; and S. Sherry and E.H. Lai for discussion. This work is funded in part by grants from the National Human Genome Research Institute (HG01720) and the SNP Consortium.

Received 7 November 2000; accepted 1 March 2001.

Gabor Marth[1], Raymond Yeh[3], Matthew Minton[2], Rachel Donaldson[2], Qun Li[2], Shenghui Duan[2], Ruth Davenport[2], Raymond D. Miller[2] & Pui-Yan Kwok[2,3]

[1]National Center for Biotechnology Information, Bethesda, Maryland, USA. [2]Division of Dermatology and [3]Department of Genetics, Washington University, St. Louis, Missouri, USA. Correspondence should be addressed to P.-Y.K. (e-mail: kwok@genetics.wustl.edu).

1. Collins, F.S., Guyer, M.S. & Chakravarti, A. *Science* **278**, 1580–1581 (1997).
2. Marshall, E. *Science* **284**, 406–407 (1999).
3. The International SNP Map Working Group *Nature* **409**, 928–933 (2001).
4. Altshuler, D. *et al. Nature* **407**, 513–516 (2000).
5. Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. & Kwok, P.-Y. *Genome Res.* **8**, 748–754 (1998).
6. Marth, G.T. *et al. Nature Genet.* **23**, 452–456 (1999).
7. Taillon-Miller, P. & Kwok, P.-Y. *Genome Res.* **9**, 499–505 (1999).
8. Rozen, S. & Skaletsky, H. *Methods Mol. Biol* **132**, 365–386 (2000).
9. Kwok, P.-Y., Carlson, C., Yager, T., Ankener, W. & Nickerson, D.A. *Genomics* **23**, 138–144 (1994).
10. Taillon-Miller, P. *et al. Nature Genet.* **25**, 324–328 (2000).

Genetic linkage of childhood atopic dermatitis to psoriasis susceptibility loci

We have carried out a genome screen for atopic dermatitis (AD) and have identified linkage to AD on chromosomes 1q21, 17q25 and 20p. These regions correspond closely with known psoriasis loci, as does a previously identified AD locus on chromosome 3q21. The results indicate that AD is influenced by genes with general effects on dermal inflammation and immunity.

AD (also known as eczema) commonly begins in infancy and early childhood, and is typified by itchy, inflamed skin. It affects 10–20% of children in Western societies and shows a strong familial aggregation[1,2]. Eighty percent of cases of AD have elevations of the total serum IgE concentration[3], and atopic mechanisms dominate current understanding of the pathogenesis of the disease[4].

We examined 148 nuclear families recruited through children with active AD (see Web Methods). The families contained 383 children and 213 sibling pairs; 254 children had physician-diagnosed AD, 153 had asthma and 139 had both. Children with AD were aged 6.9±4.4 years and 124 were male. The age of onset of disease was less than 2 years in 90% of children (geometric mean 1.5 y). We found that 51.5% of children had moderate disease and 28.6% had severe disease. The serum IgE concentration was much higher in children with AD and asthma together (geometric mean 880 IU/l; 95% CI 637–1,230 IU/l) than in children with asthma alone (mean 91; 95% CI 23–361 IU/l) or with AD alone (mean 171; 95% CI 106–277 IU/l).

We typed 385 microsatellite markers with an average marker spacing of 8.9 cM and an average information content greater than 65%. We tested four phenotypic models for linkage by non-parametric sib-pair methods. These were AD$_{ao}$ (affected subjects only), AD$_{au}$ (affected and unaffected subjects given equal weighting), asthma$_{au}$ (affected and unaffected subjects given equal weighting) and the total serum IgE analysed as a quantitative trait. We had insufficient subjects with asthma to analyse only affected sibpairs.

At the P<0.001 level, we identified linkage to AD on chromosomes 1q21 and 17q25, and linkage to asthma on 20p (Table 1). Linkage of chromosome 20p to children with both AD and asthma (χ^2=10.9, P=0.0005) was not greatly different than that to children with asthma alone, indicating that the combination of AD and asthma may correspond to a genetic subtype of disease. The total serum IgE concentration was linked to chromosome 16q–tel. Weaker evidence for linkage was seen between the total serum IgE and *D5S2115* (P=0.004) within the chromosome 5 cytokine cluster,

Table 1 • Results of linkage analysis from genome screen

		AD$_{ao}$		AD$_{au}$		Asthma$_{au}$		IgE	
Marker	Location[a]	χ^2 (LR)[b]	P[c]	χ^2 (LR)	P	χ^2 (LR)	P	χ^2 (LR)	P
D1S252	155.1	4.74	0.015	7.54	0.003	–	–	3.45	0.03
D1S498	160.7	4.00	0.02	10.95	0.0005	–	–	3.04	0.04
D1S484	173.9	–	–	5.34	0.01	–	–	–	–
D16S520	123.3	–	–	–	–	–	–	10.25	0.0007
D17S784	117.7	11.04	0.0004	5.38	0.01	–	–	–	–
D17S928	128.7	8.23	0.002	4.78	0.015	–	–	–	–
D20S889	11.0	–	–	–	–	3.86	0.02	–	–
D20S115	20.9	–	–	–	–	10.63	0.0005	–	–
D20S186	33.2	–	–	–	–	6.67	0.01	–	–

Linkages with P<0.001 are shown, together with flanking markers with P<0.05. [a]Position in cM from top of chromosome linkage group. [b]Likelihood ratio χ^2. [c]Single marker significance, unadjusted for genome-wide scan.